





Biological Databases: Defining and Building



Francis Ouellette
Director, CMMT Bioinformatics Core Facility, UBC
Vancouver, BC, Canada
francis@cmmt.ubc.ca



Current Topics in Genome Analysis
Tuesday October 31, 2000




Outline

- ◆ Bioinformatics
 - ◆ Definition
 - ◆ Information space
- ◆ GenBank
 - ◆ Format
 - ◆ Submissions and updates
- ◆ BIND
 - ◆ New database for interactions

The challenge of the information space:

Oct 18, 2000



| | |
|--------------------------------------|----------------|
| ◆ Nucleotide records | 9,102,634 |
| ◆ Nucleotides | 10,335,692,655 |
| ◆ Protein sequences | 1,183,833 |
| ◆ 3D structures | 12,863 |
| ◆ Expression data points | >20,000,000 |
| ◆ Human Unigene Clusters | 84,130 |
| ◆ Maps and Complete Genomes | 11,166 |
| ◆ Different taxonomy Nodes | 162,025 |
| ◆ dbSNP | 1,463,178 |
| ◆ Human RefGenes records | 14,133 |
| ◆ Human Contigs > 500 kb (28,515 MB) | 257 |
| ◆ PubMed records | 10,965,353 |
| ◆ OMIM records | 11,950 |

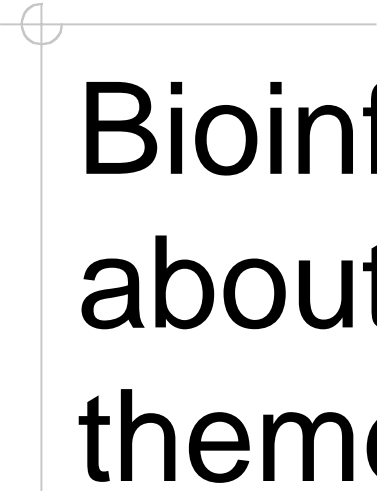

Status of the Human Genome:

(Oct 15, 2000)

| | Total sequence (kb) | Non- redundant sequence (kb) | Percentage of the Genome |
|-------------------|---------------------------|---------------------------------------|--------------------------------|
| Finished | 947,856 | 848,712 | 26.5% |
| Unfinished | 3,546,469 | 2,067,718 | 64.6% |
| Total | 4,494,325 | 2,916,430 | 91.1% |

Computational Biology (Bioinformatics)

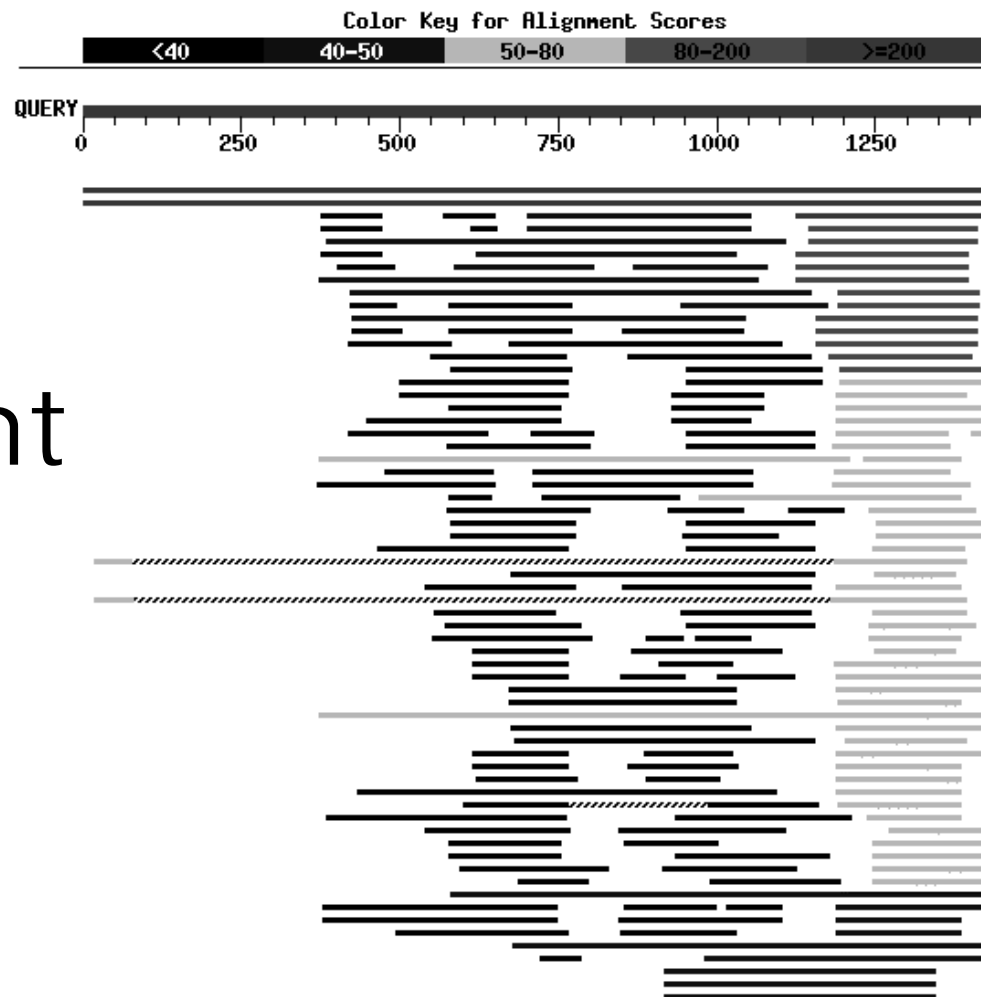
- ◆ “New” Field of Science where mathematics, computer science and biology combine together to study and interpret genomic and proteomic information.
- ◆ CB will provide the tools for fully taking advantage of the HGP (est. 2003) as well as all of the other genome projects.
- ◆ CB will position its users at the head of the pack in any race for drug target discovery as well as improving healthcare worldwide.



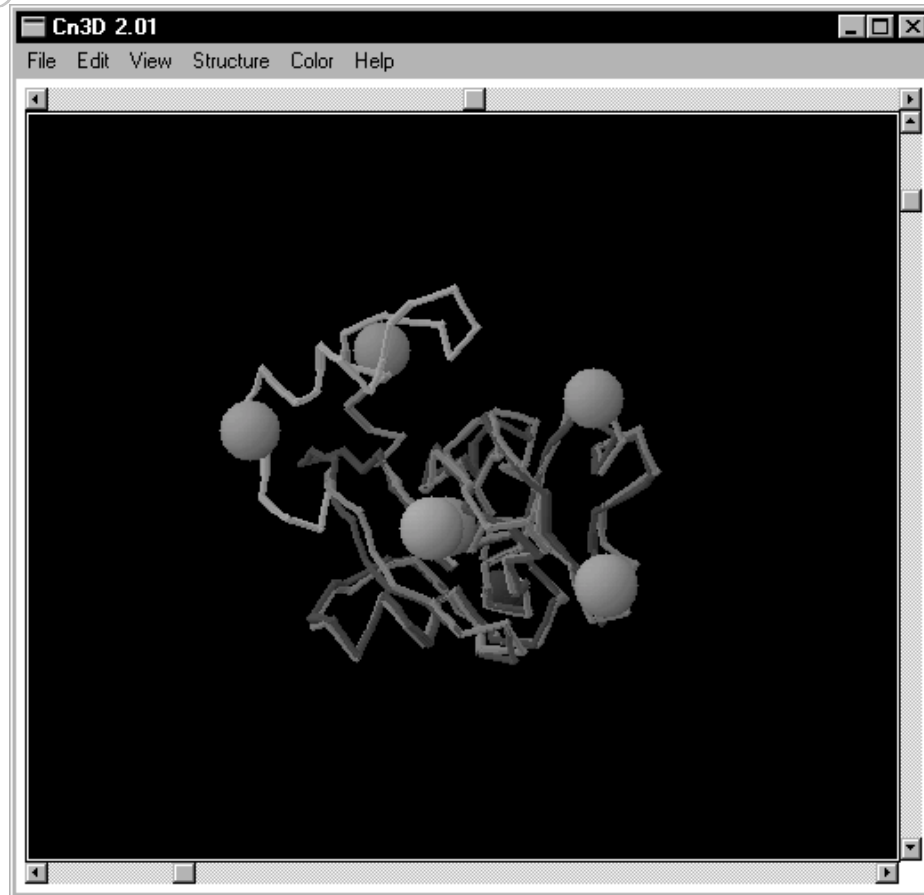
Bioinformatics is
about bringing biological
themes together with
the help of computer tools

BLAST Result

- ◆ Basic
- ◆ Local
- ◆ Alignment
- ◆ Search
- ◆ Tool



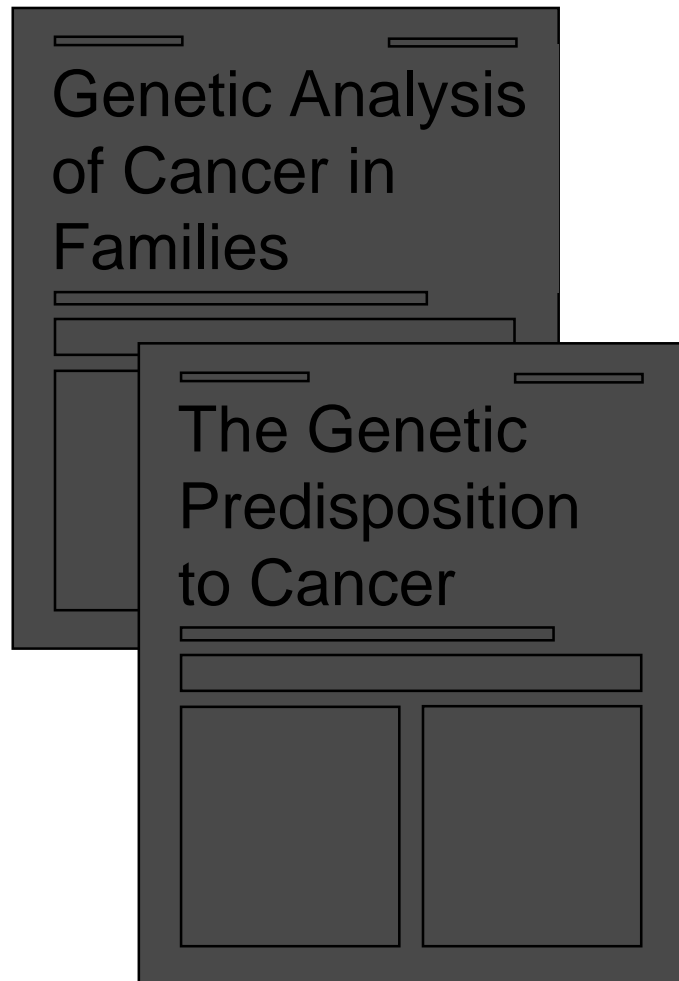
VAST result



- ◆ Vector
- ◆ Alignment
- ◆ Search
- ◆ Tool
Ferredoxin

- *Halobacterium marismortui*
- *Chlorella fusca*

PubMed Text Neighboring



- ◆ Common terms could indicate similar subject matter
- ◆ Statistical method
- ◆ Weights based on term frequencies within document and within the database as a whole
- ◆ Some terms are better than others

Micro-array analysis:

Science Jan 1 1999: 83-87

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown

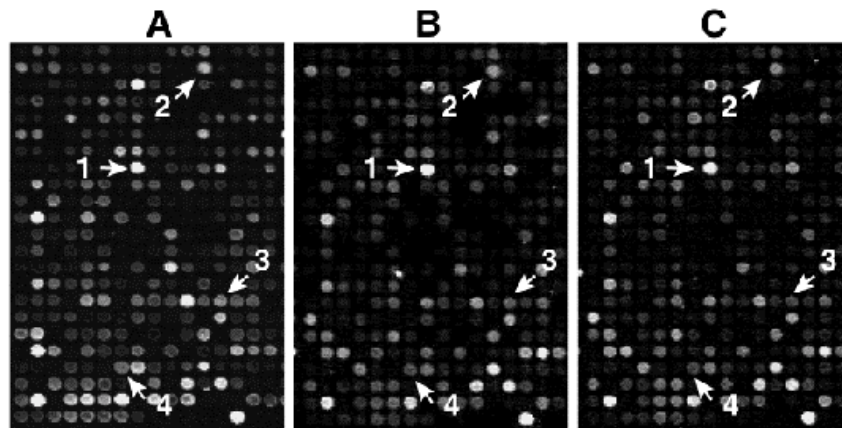


Figure 1

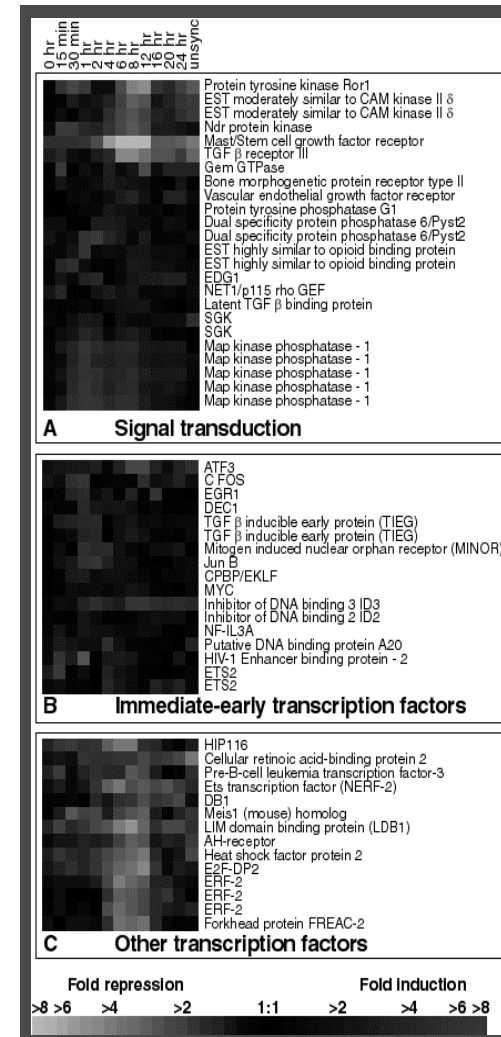


Figure 4

Databases

- ◆ Organized array of information
- ◆ Put things in, and being able to get them out again.
- ◆ Make discoveries.
- ◆ Simplify the information space by specialization.
- ◆ Resource for other databases and tools.

Database Components

- ◆ Definition and description
- ◆ Unique key
- ◆ Update version
- ◆ Links to other databases
- ◆ Documentation
- ◆ Submission/update/correction process

Information Retrieval System

- ◆ User interface
- ◆ Batch-mode
- ◆ Structured queries or SQL access
- ◆ Full-dump
- ◆ All of the data
- ◆ Documentation
- ◆ Link definitions
- ◆ User support

Cost

- ◆ Production cost
- ◆ Usage cost
- ◆ Government or academic vs industry
- ◆ Government or academic with industry
- ◆ Industry vs Industry

" ... the more closely and elegantly a model follows a real phenomenon, the more useful it is in predicting or understanding the natural phenomenon it mimics."

Jim Ostell on the "NCBI data model"

from *"Bioinformatics, a Practical Guide to the Analysis of Genes and Proteins."*, Baxevanis and Ouellette, Eds. 1998

The NCBI Data Model is defined in ASN.1

- ◆ ASN.1 is a data description language similar to a Backus-Naur Form.
- ◆ It is a formal language specifically designed to specify complex data structures in a machine, DBMS, and programming language independent manner.
- ◆ It is an international standard (ISO 8824, 8825)
- ◆ It is used by many data exchange protocols (e.g. X.400, Z39.50, WAIS).

A Bioseq defines an integer coordinate system.

- ◆ ASN.1 definition

```
Bioseq ::= SEQUENCE {  
    id          SET OF Seq-id ,  
    descr       Seq-descr          OPTIONAL,  
    inst        Seq-inst ,  
    annot       SET OF Seq-annot    OPTIONAL}
```

- ◆ The minimum required elements are an ID and the instance (e.g. length, topology, residues).



There are many classes of Bioseq

- ♦ A Bioseq may be DNA, RNA, or protein.
- ♦ A Bioseq may be represented many ways.

| | | |
|---------|-------------------------------|------------------|
| virtual | | No residues |
| raw | ———— | AGCCTTT |
| seg | ——— ——— | Parts by pointer |
| map |↑.....↑.....↑.....↑..... | Landmarks |

- ♦ A Bioseq may have a history (Seq-hist)

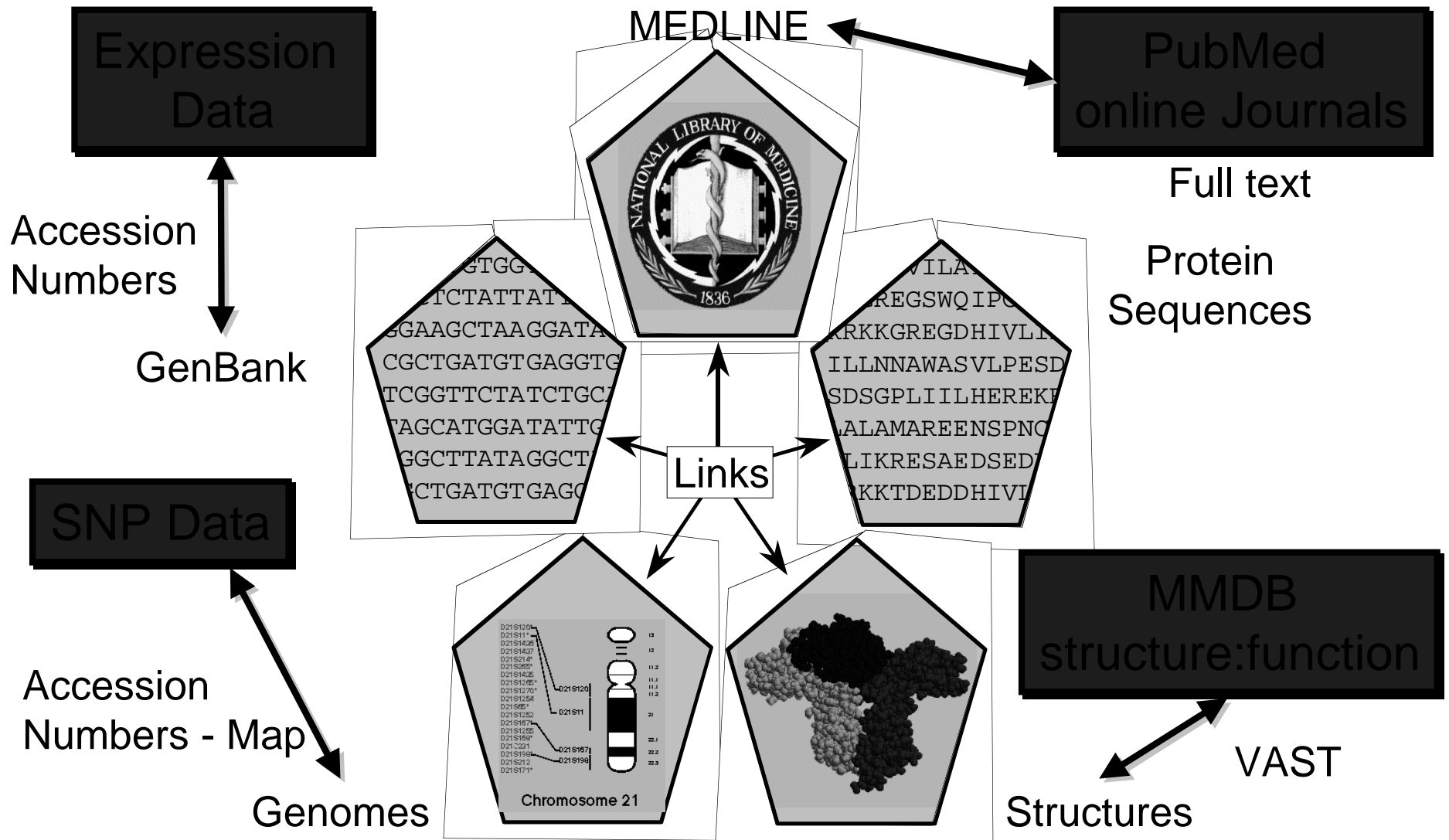


Seq-id's have different forms and usage

- ◆ Seq-id is defined as a choice of types with different forms and semantics.
- ◆ Some reflect the form and practice of the source databases or individuals.
- ◆ The NCBI “gi” is an arbitrary integer id which:
 - explicitly identifies a specific sequence
 - is stable and retrievable over time
 - has the same form over all sequence databases
 - is used to provide a history of changes to the sequence

Using the NCBI data model

CMMT



Missing?

- ◆ Full dump of the data
 - PubMed
 - Neighbors
- ◆ Documentation
- ◆ Documentation of the software to access the data
- ◆ Uniformity of the query language
- ◆ Update process for PubMed

New information space to explore

- ◆ Interactions
- ◆ Complexes
- ◆ Pathways

- New tools and databases
- New documentation
- New ways to think about the information

Primary Data

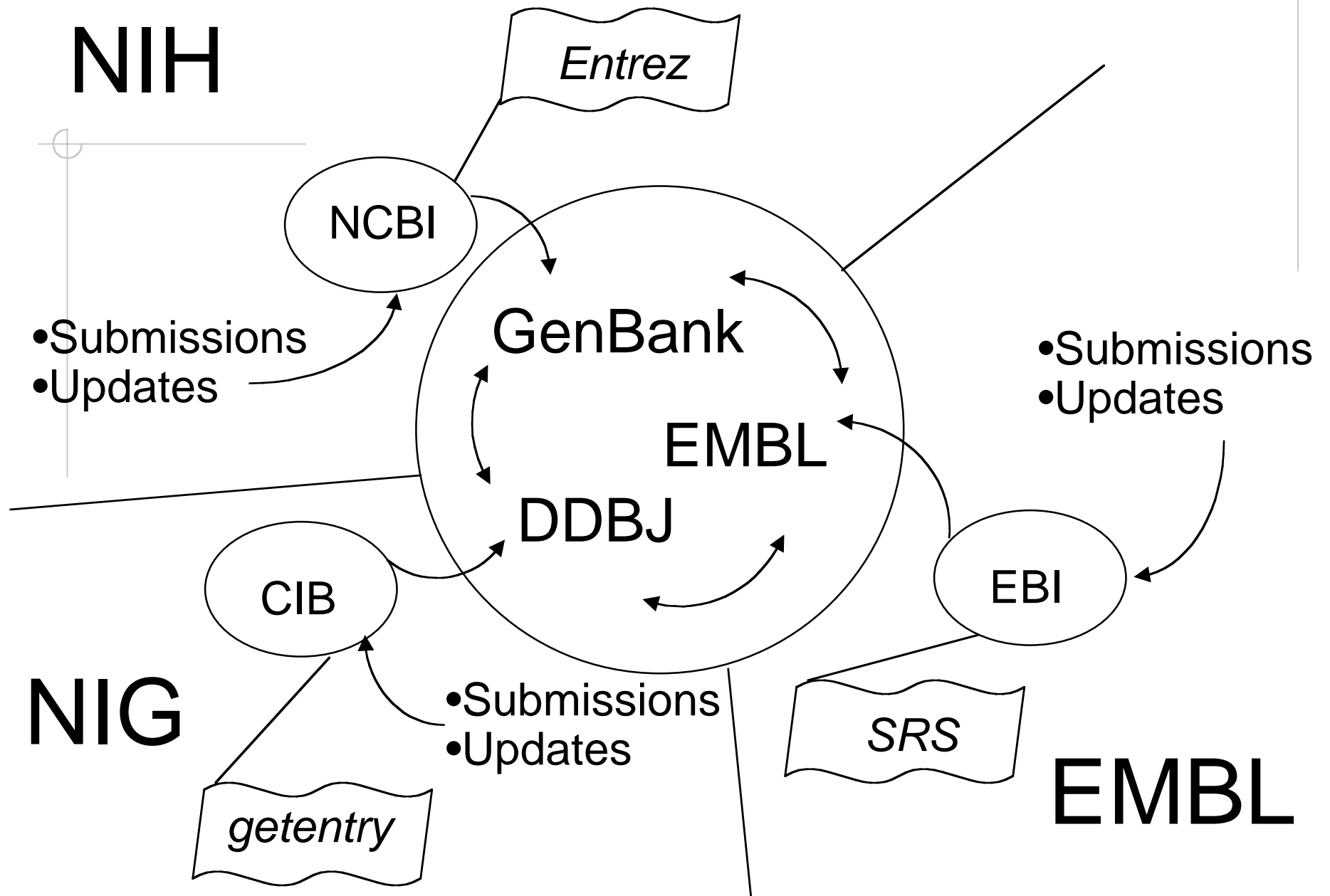
- ◆ DNA/RNA and protein sequences are the primary data for computational biology.
- ◆ In most cases protein sequences are interpreted sequences.
- ◆ Understanding the various types sequences present in GenBank is key to any interpretation in computational biology.
- ◆ Also understand that, as careful as NCBI and others are, errors do creap in, and one needs to always keep that critical eye open.

What is GenBank?

- ◆ GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

<http://www.ncbi.nlm.nih.gov/GenBank/GenbankOverview.html>

Benson *et al.*, 2000, *Nucleic Acids Res.* **28**:15-18



GenBank - Release 120 - Oct 2000

> 160,000 “species” or “terminal nodes”

9,102,634 entries or GBFF

10,335,692,655 nucleotides

- ◆ Full release of GenBank every 2 months.
- ◆ Incremental and cumulative releases: daily.
- ◆ GenBank is only available from the Internet.

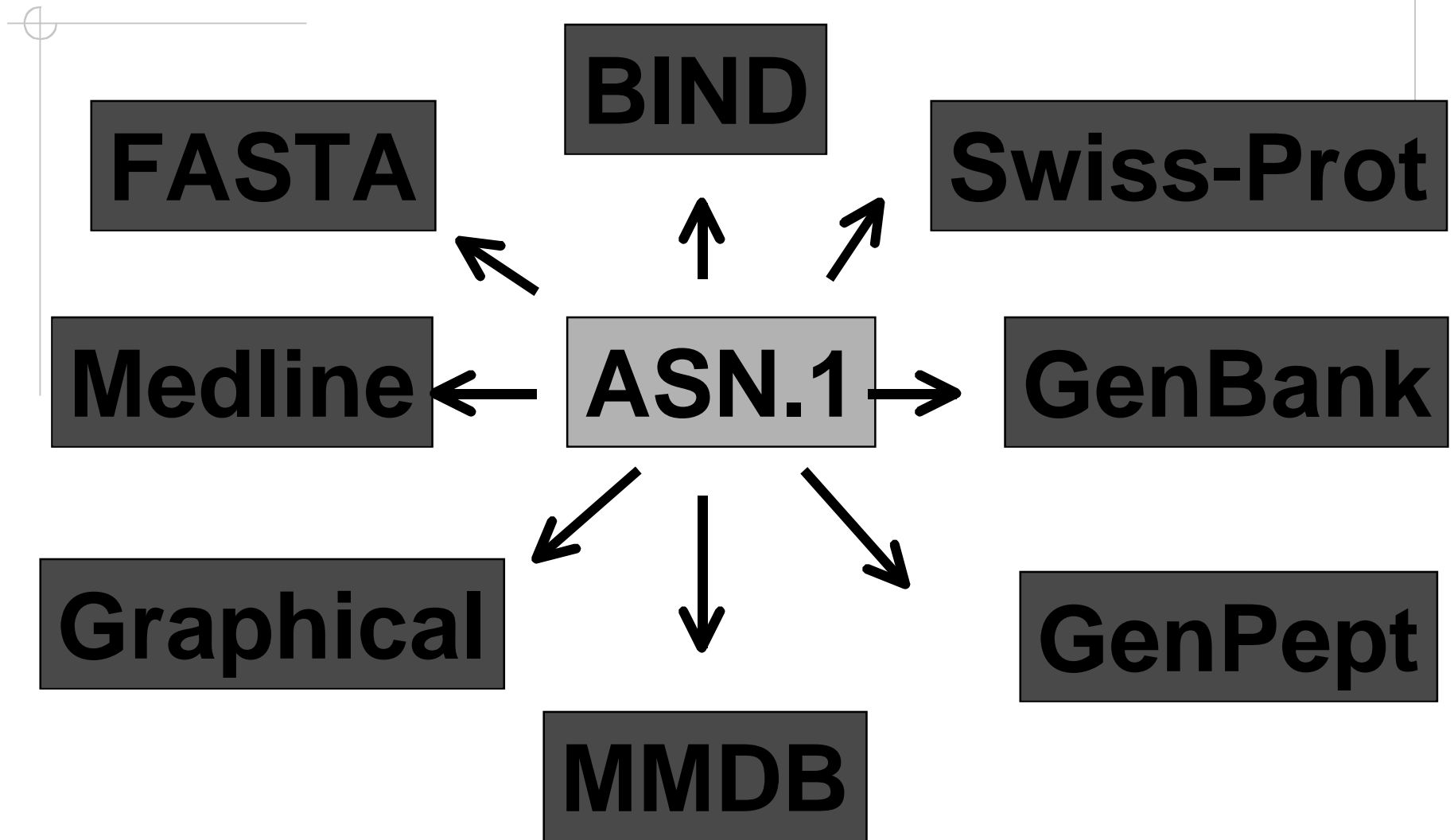
Some insights into using GenBank

- ◆ GenBank is a nucleotide-centric view of the information space.
- ◆ GenBank is a repository of all publicly available sequences. If it's not in GenBank, it might as well not be considered part of the "public domain".
- ◆ In GenBank, records are grouped for various reasons: understand this is key.
- ◆ Data in GenBank is only as good as what you put in: applying this is quite important.

GBFF and ASN.1

- ◆ GenBank data is maintained at the NCBI in the ASN.1 format.
- ◆ ASN.1 is a language that is used by computers to store, maintain, validate and show sequence information – not meant for ‘human reading’.
- ◆ The GenBank Flat File (GBFF) is one of these views (report) you can generate from ASN.1, but has taken a life of its own in the bioinformatics community.

ASN.1 as the CB lingua franca



Sample GenBank Record

LOCUS HSU40282 1789 bp mRNA PRI 21-MAY-1998
DEFINITION Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION U40282
VERSION U40282.1 GI:3150001
KEYWORDS .
SOURCE human.
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1789)
AUTHORS Hannigan,G.E., Leung-Hagesteijn,C., Fitz-Gibbon,L., Coppolino,M.G.,
Radeva,G., Filmus,J., Bell,J.C. and Dedhar,S.
TITLE Regulation of cell adhesion and anchorage-dependent growth by a new
beta 1-integrin-linked protein kinase
JOURNAL Nature 379 (6560), 91-96 (1996)
MEDLINE 96135142
REFERENCE 2 (bases 1 to 1789)
AUTHORS Dedhar,S. and Hannigan,G.E.
TITLE Direct Submission
JOURNAL Submitted (07-NOV-1995) Shoukat Dedhar, Cancer Biology Research,
Sunnybrook Health Science Centre and University of Toronto, 2075
Bayview Avenue, North York, Ont. M4N 3M5, Canada
REFERENCE 3 (bases 1 to 1789)
AUTHORS Dedhar,S. and Hannigan,G.E.
TITLE Direct Submission
JOURNAL Submitted (21-MAY-1998) Shoukat Dedhar, Cancer Biology Research,
Sunnybrook Health Science Centre and University of Toronto, 2075
Bayview Avenue, North York, Ont. M4N 3M5, Canada
REMARK Sequence update by submitter
COMMENT On May 21, 1998 this sequence version replaced gi:2648173.

GenBank Flat File (GBFF)

```
LOCUS       M25291             1803 bp     mRNA             ROD             29-AUG-1997
DEFINITION  Mouse neuroblastoma and rat glioma hybridoma cell line NG108-15
            cell TA20 mRNA, complete cds.
ACCESSION   D25291
VERSION     GI1850791
KEYWORDS    neurite extension activity; growth arrest; TA20.
SOURCE      Murinae gen. sp. mouse neuroblastma-rat glioma hybridoma
            cell_line:NG108-15 cDNA to mRNA.
ORGANISM    Murinae gen. sp.
            Eukaryotae; mitochondrion eukaryotes; Metazoa; Chordata;
            Vertebrata; Mammalia; Eutheria; Rodentia; Sciurognathi; Muridae;
            Murinae.
REFERENCE   1 (sites)
AUTHORS     Tohda,C., Nagai,S., Tohda,M. and Nomura,Y.
TITLE       A novel factor, TA20, involved in neuronal differentiation: cDNA
            cloning and expression
JOURNAL     Neurosci. Res. 23 (1), 21-27 (1995)
MEDLINE     96064354
REFERENCE   3 (bases 1 to 1803)
AUTHORS     Tohda, C.
TITLE       Direct Submission
JOURNAL     Submitted (18-NOV-1993) to the DDBJ/EMBL/GenBank databases. Chihiro
            Tohda, Toyama Medical and Pharmaceutical University, Research
            Institute for Wakan-yaku, Analytical Research Center for
            Ethnomedicines; 2650 Sugitani, Toyama, Toyama 930-01, Japan
            (E-mail:CHIHIROW@toyama-mpu.ac.jp, Tel:+81-764-34-2281(ex.2841),
            Fax:+81-764-34-5057)
COMMENT     On Feb 26, 1997 this sequence version replaced gi:793764.
FEATURES             Location/Qualifiers
     source          1..1803
                     /organism="Murinae gen. sp."
                     /note="source origin of sequence, either mouse or rat, has
                     not been identified"
                     /db_xref="taxon:39108"
                     /cell_line="NG108-15"
                     /cell_type="mouse neuroblastma-rat glioma hybridoma"
     misc_signal     156..163
                     /note="AP-2 binding site"
     GC_signal       647..655
                     /note="Sp1 binding site"
     TATA_signal     694..701
     gene            748..1311
                     /gene="TA20"
     CDS             748..1311
                     /gene="TA20"
                     /function="neurite extension activity and growth arrest
                     effect"
                     /codon_start=1
                     /db_xref="PID:d1005516"
                     /db_xref="PID:g793765"
                     /translation="MKGCLWVSRSLPNSPNHYRSPLSHTLHIRYNNLSLFINTHLSRR
                     KLRVTNPIYTRKRSINLIFYLLIPSCRTLLIWIIVYRNLIKHWSTSTVSHSHSIYRL
                     RFNMKTNILLKCHSYKPPISHFIYNNFSPRNLRGLLSRQSHLOPILRFPLPLTIYY
                     RFSNRSFPLPFRNRKQPNRIKLRCK"
     polyA_site      1803
BASE COUNT      507 a      458 c      311 g      527 t
ORIGIN
1  toagtttttt ttttttttt ttttttttt ttttttttt ttttttttt ttgtattcatg
61  tccgtttaca ttgttaagt tcacaggcct cagtcaacac aattggactg ctacaggaat
121  cctccttggt gaccgcagta tacttgccct atgaacccaa gccacctatg gctaggtagg
181  agaagctcaa ctgtagggtc gactttggaa gagaatgccat atggctgtat cgacatttca
241  catggtggac ctctgcccag agtcagcagg ccagggttct tcttcgggcg tgctccctca
301  ctgcttgact ctgctcagc ggttcacatc tgtggcgaga cgttatctgc atttgcttc
361  catctcgac  ggcattgctc ccatctagct gagaaggagc agagcctggt tctctagggc
421  gtttcattg  ggcctggtg acaatccaaa agatgaggcg tcacaacac  agaatcagaa
481  gcccacgct  attgtaaaa acactttctg gtgggaatga atggtacagg ggcgtttcag
541  gacaagagac agctttttct tcactccatc gagaacgctc gcaatcactg ttccgaagag
601  gagagatca  gaatacagc gtatggcctc gcagatgcgc cgaagagagc cagagcccat
661  ggaagcagaa agacgaaaaa cacaccattt attaaaatt attaaccaat catctctga
721  cctacactgc ccatccaaca ttctcatcag atgaacttt ggcctcctc taggagctcg
781  cctaatagtc caatcatta  caggtctttt cttagccata cactacacat cagatacaat
841  aacagcttt  tatcagtaa acacacattg tcagacagta aatacaggtt gactaatccg
901  atatacaac  gcaaacagg cctcaattt ttattttg  ttattccttc atgtcgagg
961  aggettatat tatggatcat atacatttat agaaacctga aacattggag tacttctact
1021  gttcgagctc atagccacag catttatagg ctactcctt ccatgaggac aaatcattt
1081  ctgaggtgcc acagtattta caaacctcct atcagccatc coatatattg gaacaaacct
1141  agtgaatga  atttgagggg gcttctcagt agacaaagcc acctgaacc  gattcttcgc
1201  ttccaacttc attctaacat ttattatcgc ggccttaga  atcgttacc  tctcttctt
1261  ccacgaacaa ggcatacaaa acccaacagg attaaacta gatcgagata aaattccatt
1321  tcacccctac tatacatcaa agatatccta ggtatcctaa tcatattctt aatttcata
1381  accactgatt tatttttccc agacatacta ggaacccagc acaactacat accagctaat
1441  coactaaca  ccccccacca tatcaaaccc gaatgatatt tccatttgc  atacgcattt
1501  ctacgctcaa tcccaataa  actaggaggt gcttagcctt taattctatc tatcttaatt
1561  ttacgctcaa tacttttctt tcaactccta aagcaacgaa gctaatatt  cgcgccaatc
1621  acacaaattt tgtactgaat cctagtagcc aacctactta tctaacctg  aattgggggc
1681  caaccagtag acaaccattt attatcattg gccactagc  ctccattcca tacttctcaa
1741  tcatcttaat tcttatacca atctcaggaa ttatcgagaa caaatacta  aaattatatt
1801  cat
```

Header

Features

Sequence

Sample GenBank Record

LOCUS HSU40282 1789 bp mRNA PRI 21-MAY-1998
DEFINITION Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION U40282
VERSION U40282.1 GI:3150001
KEYWORDS .
SOURCE human.
ORGANISM Homo sapiens
Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE 1 (bases 1 to 1789)
AUTHORS Hannigan,G.E., Leung-Hagesteijn,C., Fitz-Gibbon,L., Coppolino,M.G.,
Radeva,G., Filmus,J., Bell,J.C. and Dedhar,S.
TITLE Regulation of cell adhesion and anchorage-dependent growth by a new
beta 1-integrin-linked protein kinase
JOURNAL Nature 379 (6560), 91-96 (1996)
MEDLINE 96135142
REFERENCE 2 (bases 1 to 1789)
AUTHORS Dedhar,S. and Hannigan,G.E.
TITLE Direct Submission
JOURNAL Submitted (07-NOV-1995) Shoukat Dedhar, Cancer Biology Research,
Sunnybrook Health Science Centre and University of Toronto, 2075
Bayview Avenue, North York, Ont. M4N 3M5, Canada

Sample GenBank Record

| FEATURES | Location/Qualifiers |
|----------|--|
| source | 1..1789 /organism="Homo sapiens" /db_xref="taxon:9606" /chromosome="11" /map="11p15" /cell_line="HeLa" |
| gene | 1..1789 /gene="ILK" |
| CDS | 157..1515 /gene="ILK" /note="protein serine/threonine kinase" /codon_start=1 /product="integrin-linked kinase" /protein_id="AAC16892.1" /db_xref="GI:3150002" /translation="MDDIFTQCREGNAVAVRLWLDNTENDLNQGDDHGFSPLHWACRE ADLSNMEIGMKVALEGLRPTIPPGISPHVCKLMKICMNEDPAKRPKFDMIVPILEKMQ DK" |

BASE COUNT 443 a 488 c 480 g 378 t

ORIGIN

1 gaattcatct gtcgactgct accacgggag ttccccggag aaggatcctg cagccccgagt

1741 ccgcctgtca caataaagtt tattatgaaa aaaaaaaaaa aaaaaaaaaa

//

LOCUS, Accession, Accession.version & gi

| | | | | | |
|------------|---|---------|------|-----|-------------|
| LOCUS | HSU40282 | 1786 bp | mRNA | PRI | 28-NOV-1997 |
| DEFINITION | Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds. | | | | |
| ACCESSION | U40282 | | | | |
| VERSION: | U40282.1 GI: 3150001 | | | | |

LOCUS: HSU40282
ACCESSION: U40282
Nucleotide gi: 3150001
VERSION: U40282.1 GI: 3150001
Protein gi: 3150002
protein_id: AAC16892.1

| | |
|-----|---|
| CDS | 157..1515 /gene="ILK" /note="protein serine/threonine kinase" /codon_start=1 /product="integrin-linked kinase" /db_xref="GI:3150002" /protein_id="AAC16892.1" |
|-----|---|

LOCUS, Accession, Accession.version & gi

LOCUS: Unique string of 10 letters and numbers in the database. Not maintained amongst databases, and is therefore a poor sequence identifier.

ACCESSION: A unique identifier to that record, citable entity; does not change when record is updated. A good record identifier, ideal for citation in publication.

Nucleotide gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

Accession.version: New system (expected late 1998) where the accession and version play the same function as the accession and gi number.

Protein gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

protein_id: new identifier which will have the same structure and function as the nucleotide Accession and version numbers.

GenBank - Release 120

| <u>GB division</u> | <u>Nucleotides</u> |
|--------------------|--------------------|
| Organismal | 2,300,497,789 |
| EST | 2,451,695,768 |
| HTG | 4,402,496,751 |
| GSS | 1,051,117,888 |
| PAT | 72,022,274 |
| STS | 51,227,345 |

GenBank Organismal divisions:

PRI - Primate

ROD - Rodent

MAM - Mammalian

VRT - Vertebrate

INV - Invertebrate

PLN - Plant

BCT - Bacterial

RNA - Structural

VRL - Viral

PHG - Phage

SYN - Synthetic

UNA - Unannotated

Functional Divisions

PAT - Patent

EST - Expressed Sequence Tags

STS - Sequence Tagged Sites

GSS - Genome Survey Sequences

HTG - High Throughput Genome

EST: Expressed sequence Tag

Expressed sequence Tags are short (300-500 bp) single reads from mRNA (cDNA) which are produced in large numbers. They represent a snapshot of what is expressed in a given tissue, and developmental stage.

Also see: <http://www.ncbi.nlm.nih.gov/dbEST/>
<http://www.ncbi.nlm.nih.gov/UniGene/>

STS: Sequenced Tagged Sites

Sequenced Tagged Sites, are operationally unique sequence that identifies the combination of primer pairs used in a PCR assay that generate a mapping reagent which maps to a single position within the genome.

Also see: <http://www.ncbi.nlm.nih.gov/dbSTS/>
<http://www.ncbi.nlm.nih.gov/genemap99/>

GSS: Genome Survey Sequences

Genome Survey Sequences are similar in nature to the ESTs, except that its Sequences are genomic in origin, rather than cDNA (mRNA).

The GSS division contains:

- random "single pass read" genome survey Sequences.
- single pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be)
- exon trapped genomic Sequences
- Alu PCR Sequences

Also see: <http://www.ncbi.nlm.nih.gov/dbGSS/>

HTG: High Throughput Genome

High Throughput Genome Sequences are unfinished genome sequencing efforts records. Unfinished records have gaps in the nucleotides sequence, low accuracy, and no annotations on the records.

Also see: <http://www.ncbi.nlm.nih.gov/HTGS/>
Ouellette and Boguski (1997) Genome Res. **7**:952-955

HTGS in GenBank

phase 0 → ←←←←←←← ← → ← → ← HTG
Acc = AC000003 gi = 116790

phase 1 → ← → → ← HTG
Acc = AC000003 gi = 1556454

phase 2 → → → HTG
Acc = AC000003 gi = 2182283

phase 3 → PRI
Acc = AC000003 gi = 2204282

40,000 to 120,000 bp

HTG: phase 1 (DRAFT)

```
LOCUS      HSAC000003 120000 bp      DNA      HTG      20-SEP-1996
DEFINITION *** UENCING IN PROGRESS *** Chromosome 17 genomic sequence; HTGS
            phase 1, 6 unordered pieces.
ACCESSION  AC000003
KEYWORDS   HTG; HTGS_PHASE1.
...
COMMENT    ***                                     ***
            *** WARNING: Phase 1 High Throughput Genome sequence ***
            ***                                     ***
            * This sequence is unfinished. It consists of 6 contigs for
            * which the order is not known; their order in this record is
            * arbitrary. In some cases, the exact lengths of the gaps
            * between the contigs are also unknown; these gaps are presented
            * as runs of N as a convenience only. When uencing is complete,
            * the sequence data presented in this record will be replaced
            * by a single finished sequence with the same accession number.
            *      1      22526: contig of 22526 bp in length
            *    22527      23035: gap of unknown length
            *    23036      33919: contig of 10884 bp in length
            *    33920      34427: gap of unknown length
            *    34428      61877: contig of 27450 bp in length
...
//
```

HTG: phase 3 (Finished)

LOCUS AC000003 122228 bp DNA PRI 07-OCT-1997
DEFINITION Homo sapiens chromosome 17, clone 104H12, complete sequence.
ACCESSION AC000003
NID g2204282
KEYWORDS HTG.
SOURCE human.

...

COMMENT The Staden databases, finishing information, and all chromatographic files used in the assembly of this clone are available from our anonymous ftp site.

All repeats were identified using RepeatMasker: Smit, A.F.A. & Green, P. (1996-1997)

<http://ftp.genome.washington.edu/RM/RepeatMasker.html>.

FEATURES Location/Qualifiers
source 1..122228
/organism="Homo sapiens"
/db_xref="taxon:9606"
/clone="104H12"
/clone_lib="Research Genetics/Cal Tech CITB978SK-B (plates 1-194)"
/chromosome="17"
repeat_region 261..370
/rpt_family="MLT1B"

Guiding principles

- ◆ In GenBank, records are grouped for various reasons, be it in organismal or functional divisions: understanding this is key to being able to fully exploit this database.

Why submit sequences to GenBank?

- ◆ No longer submit Sequences to Journal
- ◆ Journal scanning is no longer taking place
- ◆ Electronic format more useful and allows validations
- ◆ Sequences sent to DDBJ/EMBL/GenBank are exchanged daily
- ◆ Best way to exchange new data, and updates

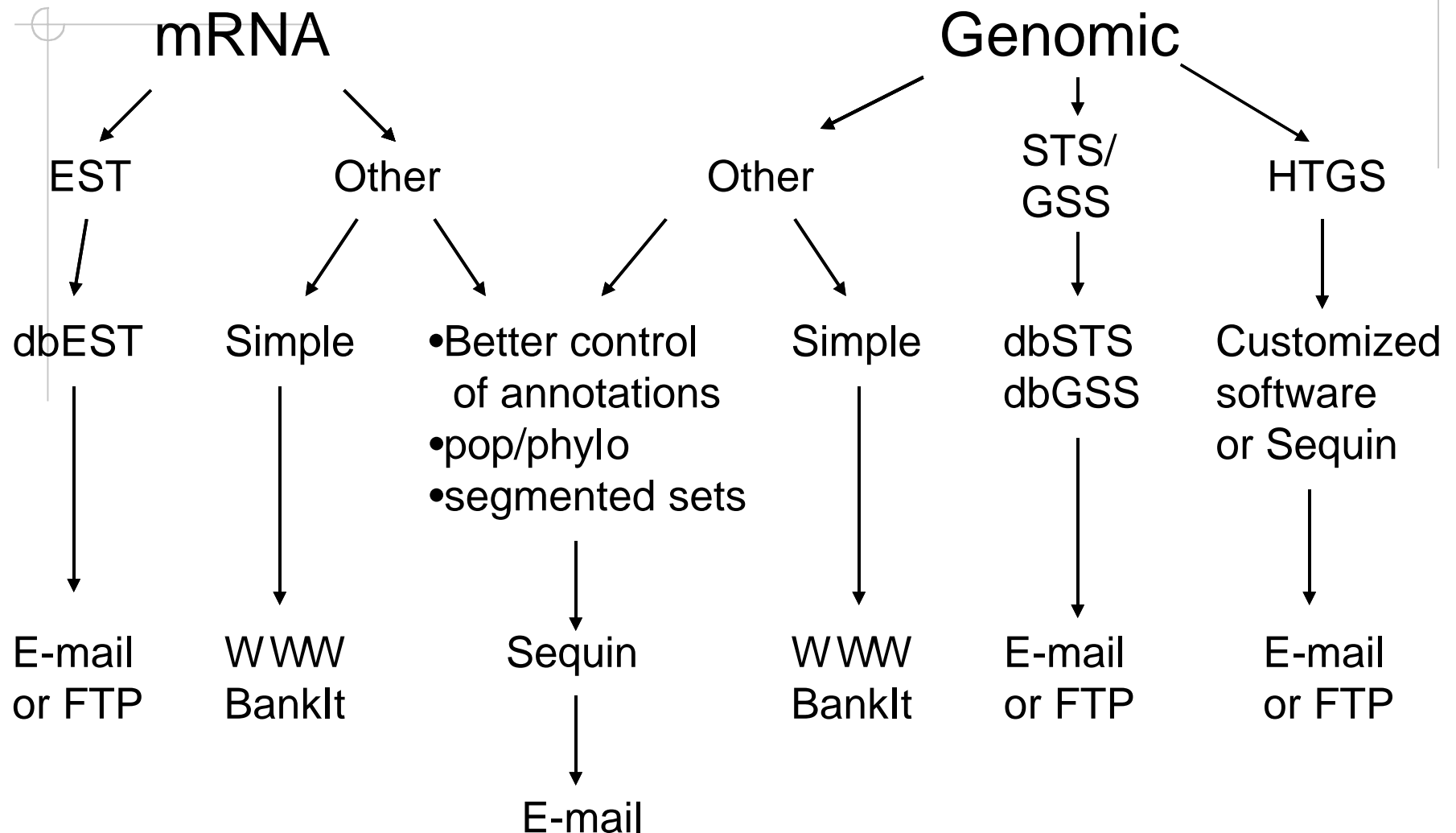
Which Tool?

- ◆ BankIt: Web based tool which is simple, easy to use, great for simple submissions, but not ideal for complicated ones.
- ◆ Sequin: Client that you need to d/I to your computer, a little harder to learn, but has great documentation, and ideal for complicated, large, multiple submissions.

Sequin

- ◆ <http://www.ncbi.nlm.nih.gov/Sequin/>
- ◆ Sequence editor for new submissions or updates
- ◆ multi-platform (Mac/PC/Unix)
- ◆ built-in validation suite
- ◆ can do:
 - segmented sets
 - pop/phylo sets
 - large records
 - different views
 - specialized editors
 - complex or simple annotations
 - BLAST and *Entrez* client

Which tool?



Where to Submit?

- ◆ Sequin files are e-mailed to:

`gb-sub@ncbi.nlm.nih.gov`

- ◆ BankIt:

<http://www.ncbi.nlm.nih.gov/BankIt/>

- ◆ EST/GSS/STS send e-mail to:

`batch-sub@ncbi.nlm.nih.gov`

- ◆ HTGS send query e-mail to:

`htgs@ncbi.nlm.nih.gov`

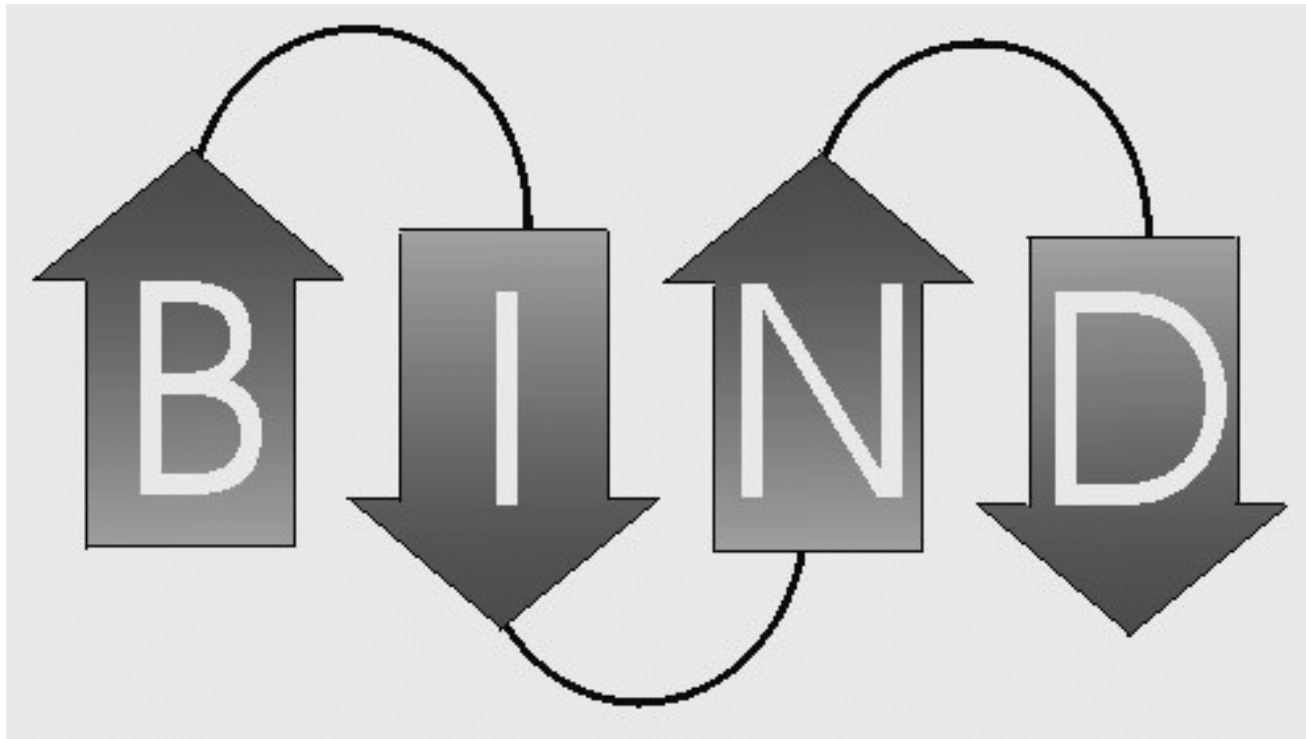
- ◆ Updates:

`updates@ncbi.nlm.nih.gov`

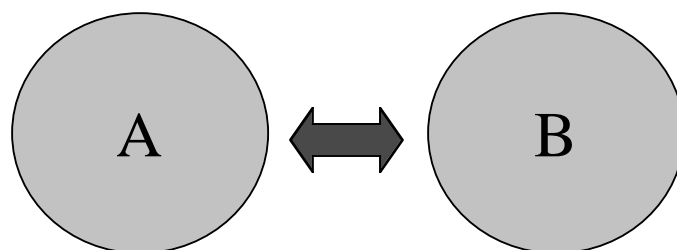
- ◆ Not sure? e-mail to:

`info@ncbi.nlm.nih.gov`

Biomolecular Interaction Network Database



A simple BIND INTERACTION record

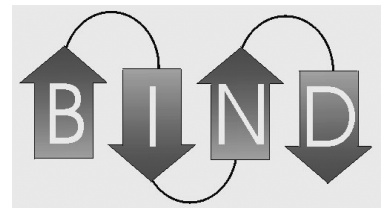


1. Short label
2. Type of molecule
3. Database identifier
4. Origin

5. Short label
6. Type of molecule
7. Database identifier
8. Origin

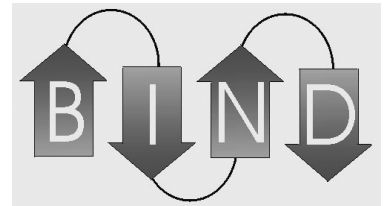
9. Publication reference

“Eventually we will have to put all these parts together ... And from this we will be able to model cells and biological processes (diseases and ‘wellness’)”



Interaction Space ...

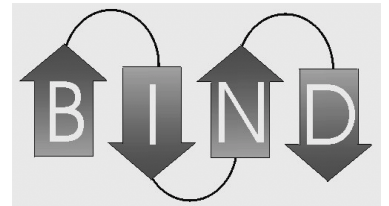
... is not available for traversal in our current Bioinformatics database paradigm.



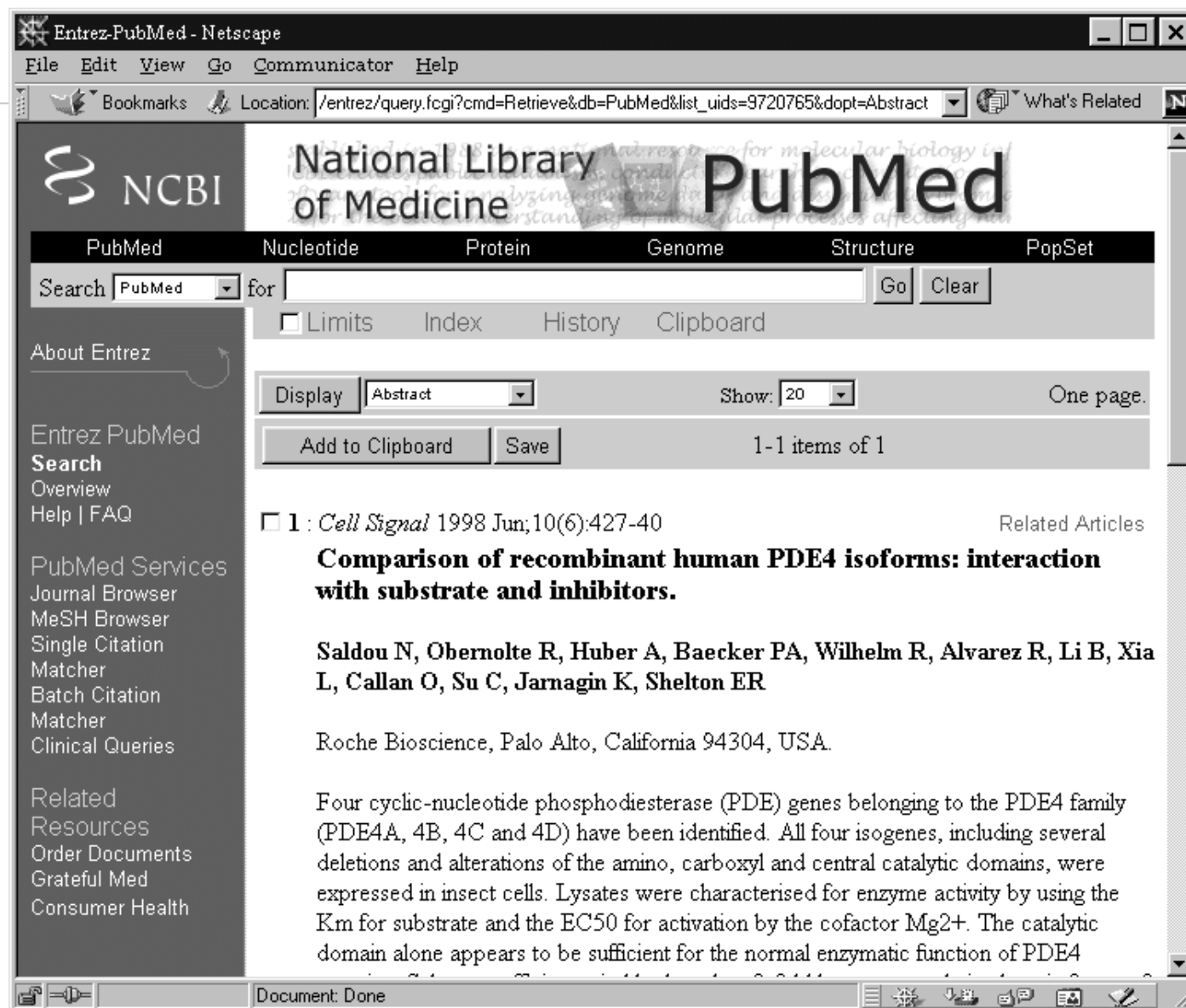
Interaction Space Query

Query:

What interacts with PDE4?

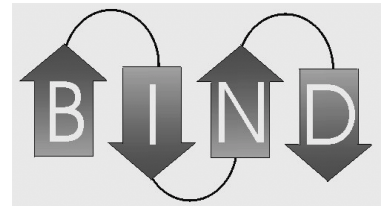


Answer: Go read the literature



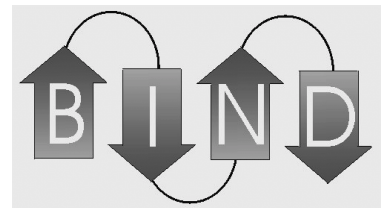
Better Answer:

- Mg²⁺
- cAMP
- Protein kinase A
- IBMX
- Trequinsin
- Rolipram
- TVX 2706
- RP 73401
- RS-25344 ...



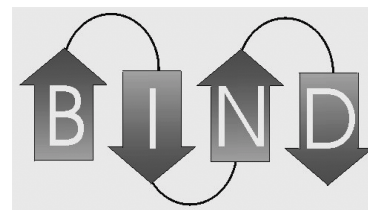
Interaction Annotations are Known

- ◆ Annotations are often about interactions
- ◆ Interactions are more and more being discovered by persons other than those who first sequence the gene.
 - ◆ Yeast two-hybrid
 - ◆ immunoprecipitation
 - ◆ reconstitution
 - ◆ optical biophysical methods
 - ◆ mass spectrometry



Biomolecular Interaction Record

- Molecule A binds molecule B (binary)
 - ◆ both things have accession number pointers to other databases OR
 - ◆ Instances of A and B as ASN.1 objects.
 - ◆ dependencies
- Things happen!
 - ◆ Chemical state change to A, B or both
- Citation
- Interaction ID (interaction accession)

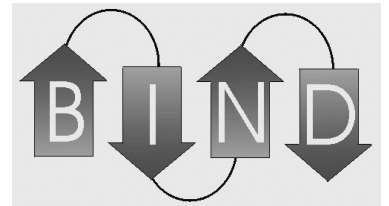


BIND

- ◆ Database contains
 - Interactions
 - Molecular complexes
 - Pathways
- ◆ Fully integrated with the NCBI biological data model

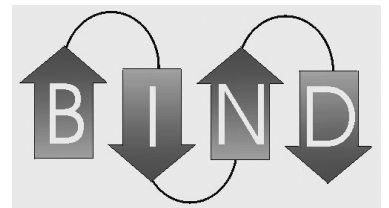
FOR MORE INFO...

<http://binddb.org>



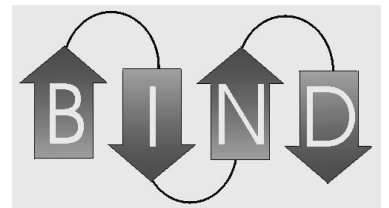
Building BIND

- ◆ Adding to the database:
 - Backfilling
 - BATCH – High throughput
 - Direct submission
- ◆ Building & developing the tools



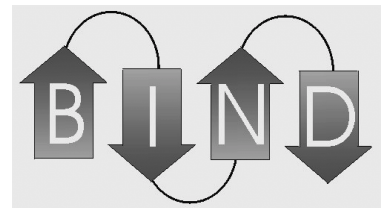
Building BIND

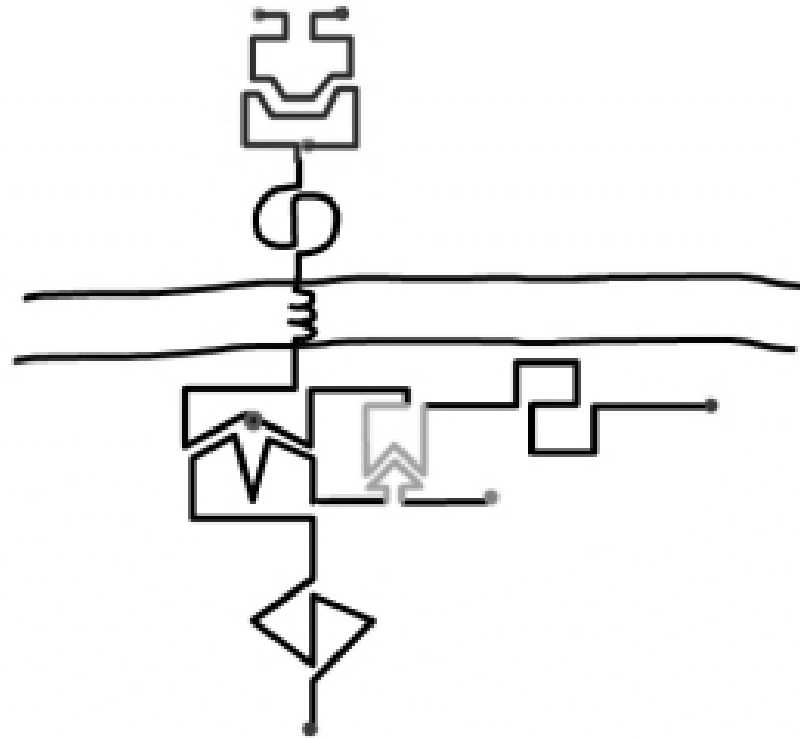
- ◆ Parallel approaches to getting data in:
 - ◆ Back-fill database with known information
 - ◆ 10 (year 1) - 50 (year 3) indexers
 - High throughput input from published datasets (*Stan Fields and the yeast interactions*).
 - Get the community to submit BIND data and obtain BIND accession numbers
- ◆ Building & developing the tools

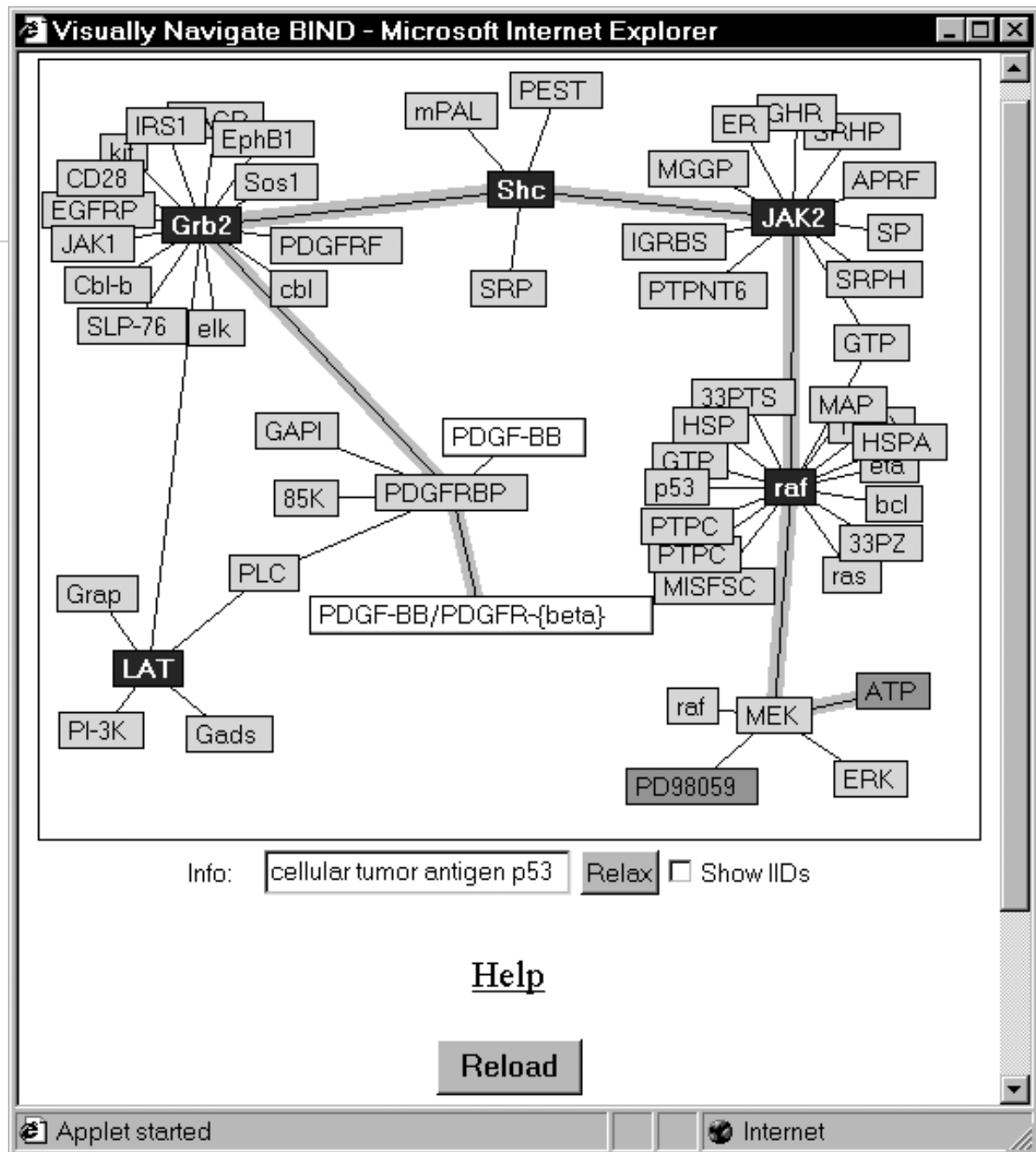


“On The Fly” Visualization Strategy

- ◆ Algorithmic generation of pathway drawings
 - ◆ User asks, “draw me a picture of HD interactions”
 - ◆ Server queries database for binding partners, assembles an image, and sends it to the user
- ◆ Define the symbolism in a creative and novel way
 - continuous line-symbols for domains
 - “mate-able”
 - we have already a library of about 500 - 1000 symbols

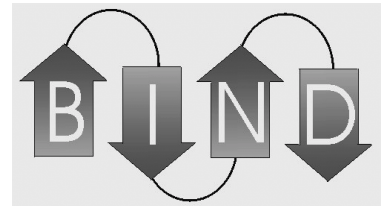






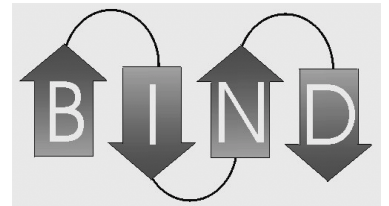
New ways to think about this data

- ◆ Where are they in the cell?
- ◆ Where do they move during pathogenesis?
- ◆ What interacts with a given protein?
- ◆ What pathways are components part of?
- ◆ What pathways are known for one organisms are also used in others?
- ◆ What are the finite interactions?
- ◆ New questions we haven't really thought about ...



Current Status:

- ◆ National collaboration:
 - Francis Ouellette
 - Christopher Hogue
 - Tony Pawson
- ◆ Database schema and prototype exists
 - Data model published:
 - ◆ Bader and Hogue, Bioinformatics
 - ◆ <http://binddb.org>
- ◆ NAR paper *in Press*, to be published in the “database issue” in January 2001.



Summary

- ◆ GenBank is a nucleotide-centric view of the information space, and is a report from the underlying ASN.1 data.
- ◆ In GenBank, records are grouped for various reasons: understand this is key to taking full advantage of this information.
- ◆ Sequin and BankIt can be used for updates and new submissions.
- ◆ Understanding the data elements in any database records is important, and allows you to take full control of the information.
- ◆ BIND is a new database that will offer new information space to allow new types of queries and discoveries.

Acknowledgments

GenBank Release Coordination

Mark Cavanaugh

GenBank Submission Coordination

Ilene Mizrahi

GenBank Annotation Staff

John Anderson, Maureen Beanan, Matthew Beyers, **Medha Bhagwat**,
Lori Black, Larry Chlumsky, **Karen Clark**, **Irene Fang**, **Michael**
Fetchko, Jeff Gilmour, Irene Kim, **Pierre Ledoux**, Richard
McVeigh, Leonie Misquitta, Michael Murphy, Cynthia
Rothblum-Oviatt, Quy Phung, **Leigh Riley**, **Susan Schafer**, Suh-suh
Wang, **Jane Weisemann**, Steven Wilhite, Sandhya Xirasagar,
Roxanne Yamashita and **Linda Yankie**

Jim Ostell & Mark Boguski

Acknowledgments

Toronto:

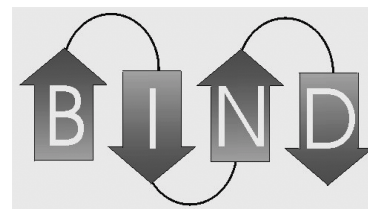
- ♦ Tony Pawson
- ♦ Christopher Hogue
 - ♦ Gary Bader
 - ♦ Ian Donaldson
 - ♦ Cheryl Wolting

Vancouver:

- ♦ Francis Ouellette
 - ♦ Patrick Franchini
 - ♦ Sohrab Shah

Ottawa:

- ♦ Joel Martin



Additional URL's from lecture:

CMMT: <http://www.cmmt.ubc.ca>
CBW : <http://www.bioinformatics.ca>
NCBI: <http://www.ncbi.nlm.nih.gov>
Genome Canada
<http://genomecanada.ca>
BIND: <http://binddb.org>
<http://bind.ca>*

* Registered but not active yet ...